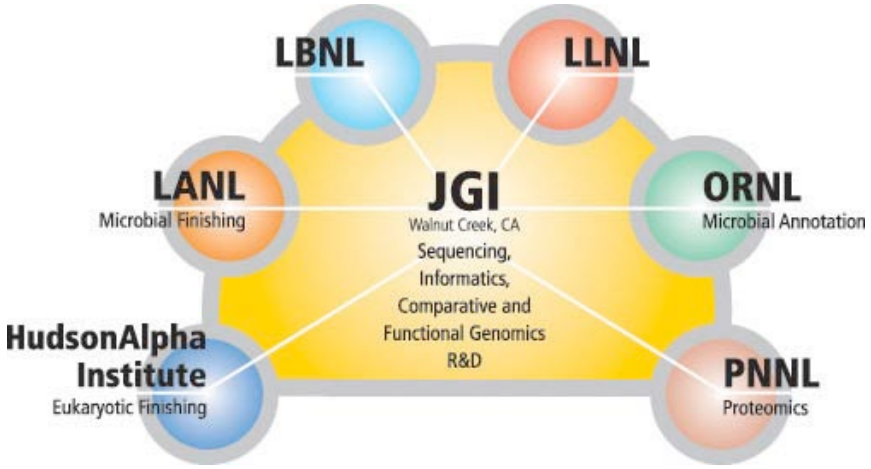
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Strategic Plan

U.S. Department of Energy
Joint Genome Institute
2010-2014

AUGUST 2009

A National Resource Enabling Large Scale Genomics and Analysis for Bioenergy and Environmental Research



The U.S. Department of Energy (DOE) Joint Genome Institute (JGI), headquartered in Walnut Creek, California, is supported by the DOE Office of Science. The DOE JGI unites the expertise of five national laboratories—Lawrence Berkeley, Lawrence Livermore, Los Alamos, Oak Ridge, and Pacific Northwest—along with the HudsonAlpha Institute for Biotechnology. Its mission is to serve as a user facility enabling the application of large-scale genomics and analysis to address DOE mission needs in bioenergy and the environment.

The United States Department of Energy (DOE) Joint Genome Institute (JGI) Five-Year Strategic Plan

2010-2014 Strategic Plan

"A National Resource Enabling Large Scale Genomics and Analysis for Bioenergy and Environmental Research."

The United States Department of Energy (DOE)
Joint Genome Institute (JGI)
Five-Year Strategic Plan

Table of Contents

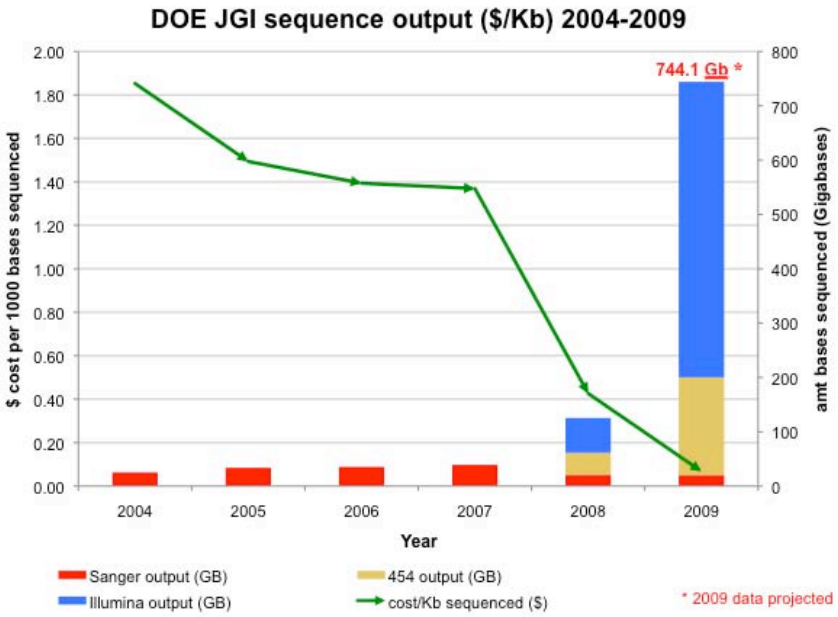
The United States Department of Energy (DOE)
Joint Genome Institute (JGI)
Five-Year Strategic Plan

Background and Overview

The DOE Joint Genome Institute (JGI) while initially created and ultimately successful in sequencing a sizeable fraction of the human genome has since focused on providing large-scale genomics capabilities to investigators working on energy and environmental problems. This has included the sequencing and analysis of the genomes of the plants, microbes and environments that participate in carbon cycling and bioenergy production. The DOE JGI represents the only user facility of its kind, dedicated to offering a variety of state of the art genomic resources and technologies to investigators working in DOE mission areas of biology.

The DOE JGI's user facility approach is based on the concept that by focusing the most advanced sequencing and analysis resources on the best peer-reviewed proposals drawn from a diverse community of scientists, the DOE JGI will catalyze creative approaches to addressing DOE mission challenges. This strategy has clearly worked, only partially reflected in the fact that the DOE JGI has played a major role in more than 50 papers published in the prestigious journals *Nature* and *Science* alone over the past three years. The involvement of a large and engaged community of users working on important energy and environmental problems has helped maximize the impact of DOE JGI science.

A seismic technological change is presently underway in the field of genomics. For the past 20-plus years, the Sanger process dominated how sequencing was done. This situation recently changed with the advent of next generation sequencers—each machine capable of generating in just a few days what the entire facility with 100 Sanger-based machines was yielding in a month. The DOE JGI is responding to these new capa-



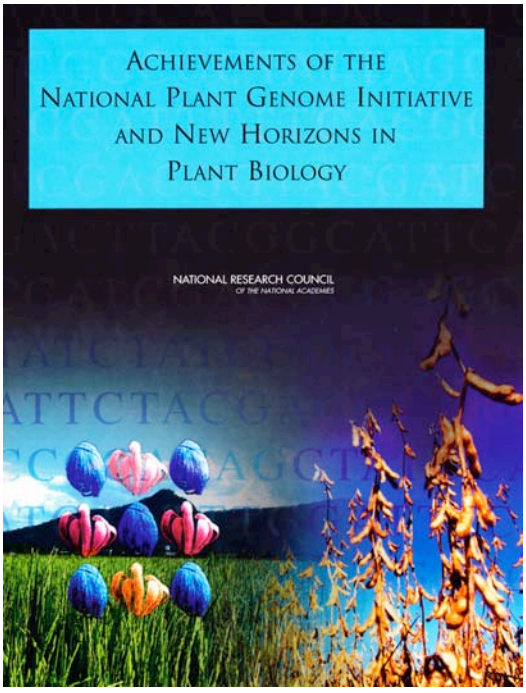
bilities as outlined in this five-year planning document, by encouraging larger scale more complex projects inconceivable with previous technologies to drive advances in bioenergy and environmental sciences.

The continued need for a large-scale genomic user facility like the DOE JGI was clearly articulated in the 2008 National Research Council reports, "*Achievements of the National Plant Genome Initiative and New Horizons in Plant Biology*," (ISBN-10: 0-309-11418-7) as well as "*The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet*" (ISBN-10: 0-309-10676-1). Both reports outline a crucial need for systematic large-scale surveys of the plant and microbial components of the biosphere as well as an increasing need for large-scale analysis capabilities to meet the challenge of converting sequence data into knowledge. The DOE JGI is extensively discussed in both reports as infrastructure that is vital to progress in these fields of major national interest. Accordingly, a major part of the DOE JGI's strategic plan for the future involves the development of a roadmap for generating and providing community access to these microbial and plant datasets, as well as the tools to analyze them. Just as the availability of the human genome has accelerated the work of biomedical researchers and contributed to the development of nearly all the drugs in today's pharmaceutical pipelines, activities of the DOE JGI microbial and plant programs proposed in this document will accelerate the work of scientists in DOE mission areas and contribute solutions to energy and environmental challenges.

As the microbial and plant genome strategic roadmap is being developed, the DOE JGI as a user facility is also evolving. With the technological changes underway, the DOE JGI will move from a central focus on "one-off" user projects coming from individual scientists and small user communities to also include targeting much larger-scale projects driven by systematic and problem-focused approaches. These "Grand Challenge" scale genomic projects will include projects many orders of magnitude larger and more complex than are presently underway. While we now, for example, characterize the metagenome of a single soil sample, in the future we expect to see projects where hundreds to thou-

sands of soil samples collected from around the US with associated "meta-data" are deeply examined resulting in insights needed for the development of diagnostic tools for improving soil management and carbon sequestration. Entire communities of scientists working in a particular field, such as biofuels feedstock improvement, biomass degradation, subsurface biogeochemistry, or climate change, will be users of this information. In addition, interactions with the DOE Bioenergy Research Centers and other DOE facilities that may follow will help shape aspects of how the DOE JGI operates as a user facility.

Based on rapidly increasing yields and diminishing costs of new sequencing technologies, the DOE JGI will maintain its prominence as a generator of sequence and, in addition, provide users an array of both pre- and post-sequencing value-added capabilities to accelerate their science. The obstacles to performing genomic-based investigations in the future will be less the generation of sequence and more the isolation of the specific material that investigators may want to sequence and the informative analysis of the sequence data after it is generated. Among the



future pre-sequencing capabilities that the DOE JGI will offer are a variety of robust ways to isolate DNA from the thousands of samples that will make up individual studies in the future, from hard-to-culture environmental microbes using single cell sorting, whole genome amplification, and high throughput culturing to capturing specific sequences of biological interest from large and complex plant genomes and environmental metagenomes. In addition, capabilities such as gene expression profiling of single organisms and mixtures of organisms as a function of natural or experimental perturbation will further facilitate DOE JGI users in accomplishing their science.

DOE Bioenergy Research Centers

A crucial post-sequencing bottleneck that the DOE JGI will increasingly focus on is the analysis of genomic data. It is already apparent that the application of genomics to problems in energy and the environment is being limited by the scientific community's inability to integrate and interpret the dramatically increasing volumes and disparate types of data being generated. The DOE JGI has established itself as a resource for the development of tools and infrastructure in response to the need of users to analyze and extract value from sequence data for solving complex problems in biology. Building on our technological capabilities, the DOE JGI will continue to expand its team of sequence analysis experts in parallel with developing the computational tools far beyond those found in small centers, to offer a unique breadth of resources to our growing community of users. In the future, the DOE JGI will also increasingly draw upon informatics and computing capabilities of the DOE National Laboratory system to facilitate storage and analysis of ever-larger genomic datasets to aid users in addressing fundamental problems.

Over the past several years, the DOE JGI has occupied a unique niche as an institute focused on applying large scale genomics to accelerate the progress of scientists working in DOE mission areas of biology. Its productivity has contributed to advances in energy and environmental biosciences, fields that have become dependent upon the microbial and plant genomes that the DOE JGI has sequenced and analyzed. With the technological advances that are occurring in genomics, the DOE JGI through the approaches and strategies described in this five year plan, is poised to have an even greater impact, contributing solutions to the nation's energy and environmental challenges.

Strategic Planning Process

The dynamic nature of the science of genomics driven by enormous and ever-changing technological advances makes it more appropriate for the DOE JGI to develop a five-year rather than a 10-year plan. The plan described below resulted from a series of comprehensive planning sessions in 2008 and 2009. Due to the importance of having genome sequence information of the major bioenergy crops and environmental microbes to accelerate the work of bioenergy and environmental researchers this plan represents a combination DOE JGI roadmap and vision for the next five years. A group of advisors, internal and external, culled from the DOE JGI Policy Board, Scientific Advisory Committee, and User Committee, as well as representatives of the three DOE Bioenergy Research Centers and several of the DOE JGI partner National Laboratories, were assigned discussion topics. A roster of the DOE JGI Five-Year Planning Group participants is listed in the Appendix.

In contemplating how the DOE JGI needs to respond to the changing landscape of genomics, it is essential to focus not just on the technology but even more importantly to align the Institute's long-term goals with the broader DOE goals in the area of biology. These include directing genomics and computational resources to:

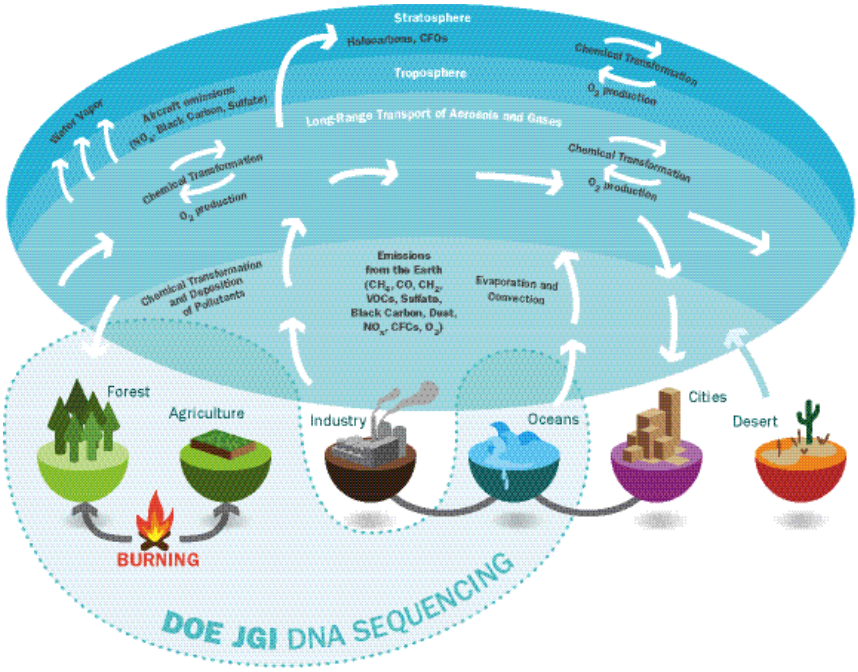
- Advance understanding of the basic biological processes underlying production of improved biofuels and associated with other energy technologies;
- Characterize the biotic factors involved in the global cycles of carbon, nitrogen and other elements;
- Characterize microbial community contributions to subsurface biogeochemistry and environmental contaminant transport and fate.

In the context of these defined goals, we focused our planning efforts on how the genomics of microbes, communities of microbes, and plants could be brought to bear on accomplishing these critical challenges.

I. Microbial Genomics & Metagenomics Program Goals

(In this document the term "microbes" refers to both prokaryotic—bacterial and archaeal—and eukaryotic [e.g., algae and fungi] microbes.)

To plan a path for the Microbial Genomics Program that is aligned with the overall goals of the DOE JGI, it is important to assess the current state of microbial genomics and metagenomics. Although there have been many advances in both areas over the last 5-10 years, it is clear that the microbial world has barely been sampled by genome sequencing and that there is an enormous amount of sequencing left to do to help characterize the "dark matter" of the microbial world. Genomic studies of this largely unstudied component of the microbial world are motivated by the crucial need for this information to facilitate the deciphering of the genetic determinants and organisms central to advancing DOE's mission in energy and the environment. This is illustrated



by the fact that while investigators at the DOE Bioenergy Research Centers and elsewhere are looking at the metagenomes of communities of microbes responsible for deconstructing plant material into fermentable sugars, their work is stymied by an inability to recognize the genes participating in the deconstruction process. The availability of a catalog or "encyclopedia" of environmental microbes, as is being proposed, will dramatically improve our ability to annotate genes from these communities and link them to function.

As the microbial program plans its next phase, it is critical to recognize that the DOE JGI has many resources currently available that can be applied to microbial projects. These resources include both the talent of the DOE JGI's workforce including at its partner labs, as well as instrumentation and computational assets optimized for production sequencing, genome finishing, and analyzing genomes and metagenomes. Coupled with these resources is a proven project management team capable of coordinating large projects and their associated diverse communities of collaborators. All of these assets will be brought to bear over the next five years to enable DOE JGI to accomplish the goals of the microbial program.

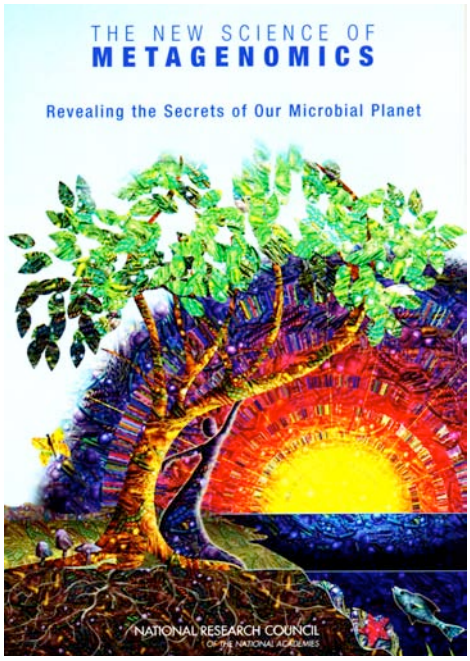
With the these resources in mind, the following targets have been identified to help achieve the goals outlined above and lead to long-term benefits for DOE mission science:

- Develop the methodology to generate routine multi-dimensional snapshots of complex environmental microbial communities, which in turn will expand our understanding of natural and contaminated environments;
- Elucidate the microbial interactions that influence global cycles of key elements such as carbon, nitrogen, and sulfur which, in turn, will facilitate our understanding of diverse natural environments;
- Establish reference genomes for all "family" level groups of microbes;
- Launch an effort to identify microbes that can be used to enhance the growth and pest resistance of major biofuel plants.

Ia. Coordinate large-scale sequencing of isolates

With the availability of more than 1000 completely sequenced bacterial and archaeal genomes, it might seem that we have come to the end of the sequencing era. While it is certainly true that making use of genome data (e.g., though functional genomic studies) is critically important, as sequencing has becoming cheaper and faster it is clear that the value one gets for sequencing a microbial genome still far outstrips the cost. The spread of sequencing has mostly led to the proliferation of small-scale projects focused on organisms already being studied in specific laboratories. At the DOE JGI, however, a new possibility has arisen – the possibility of coordinated phylogenetically targeted "encyclopedia" projects to sequence the genomes of hundreds or even thousands of isolates, at a scale beyond the capabilities of smaller facilities that would generate information crucial to the decoding of all microbial genomes.

Despite the large number of microbial genomes that have been sequenced, most come from a limited number of phyla including a marked over-representation of human pathogens. Accordingly the bacterial genomes that have been sequenced represent a very narrow subset of the phylogenetic diversity of bacteria and archea known to exist on earth. It is a bit like having a dictionary with most of the entries starting with only 3-4 letters of the 26 letter alphabet and trying to use it to decipher a text written in a foreign language. It is clear that this bias is severely hampering our ability to discover novel processes of importance to DOE mission areas in the genomes of microbes. The goal of the "encyclopedia" projects that are proposed in this plan is to systematically sequence the genomes of organisms from phyla that have not previously been sampled.



The DOE JGI has already begun conducting a pilot of one such effort, a coordinated project to sequence phylogenetic novel genomes of bacteria and archaea. This "Genomic Encyclopedia of Bacteria and Archaea"(GEBA) was selected as a large-scale project at the DOE JGI because of the clear benefits it could provide in the decoding of DOE relevant microbial genomes and environments. Although we have only completed a pilot project in this area, the data and analyses are already proving invaluable to many areas of microbial genomics and metagenomics. For example, analysis of the data from the nascent "encyclopedia" project shows that this targeted phylogenetic sampling strategy dramatically increases our ability to discover new genes and gene families. The data also clearly reveals that this filling in of the dark matter of the biological universe improves our ability to annotate metagenomic data as well as individual microbial genomes of DOE interest.

The DOE through this pilot GEBA project is learning how to carry out large-scale projects requiring the development of methods and collaborations for key rate limiting steps such as growth of organisms, acquisition of DNA, and fully-automated annotation of genomes. GEBA

has so far focused on bacterial and archaeal genomes from cultured organisms but in the next five years we will initiate other "genomic encyclopedia" type projects, ranging from GEBA surveys of unculturable microbes to focusing on eukaryotic microbes beginning with fungi, due to their central role in plant growth, biomass deconstruction and biofuel production.

Ib. Target large-scale metagenomics projects for specific types of microbial communities

In the last five years the DOE JGI has established itself as a world leader in metagenomic sequencing and analysis, applying this emerging discipline to address many key areas of interest to the DOE. In the future, the DOE JGI will carry out coordinated large-scale metagenomics projects focusing on specific environments and communities. Such large-scale "terabase" metagenomics projects are of profound importance and are one of the key recommendations of the NAS report on "*The New Science of Metagenomics*." These projects will include many components, such as isolating DNA and sequencing of thousands of "micro" environments within a single study site, deep RNA sequencing, environmental functional genomics, single cell studies, culturing and characterization of isolates, and the development of new informatics tools. Presently our view of environmental metagenomics can be likened to take a genetic snapshot of a particular environment while in the future we will want to generate the equivalent of genetic moving pictures to derive deeper insights into DOE relevant processes.

Ic. Produce high-quality finished genomes

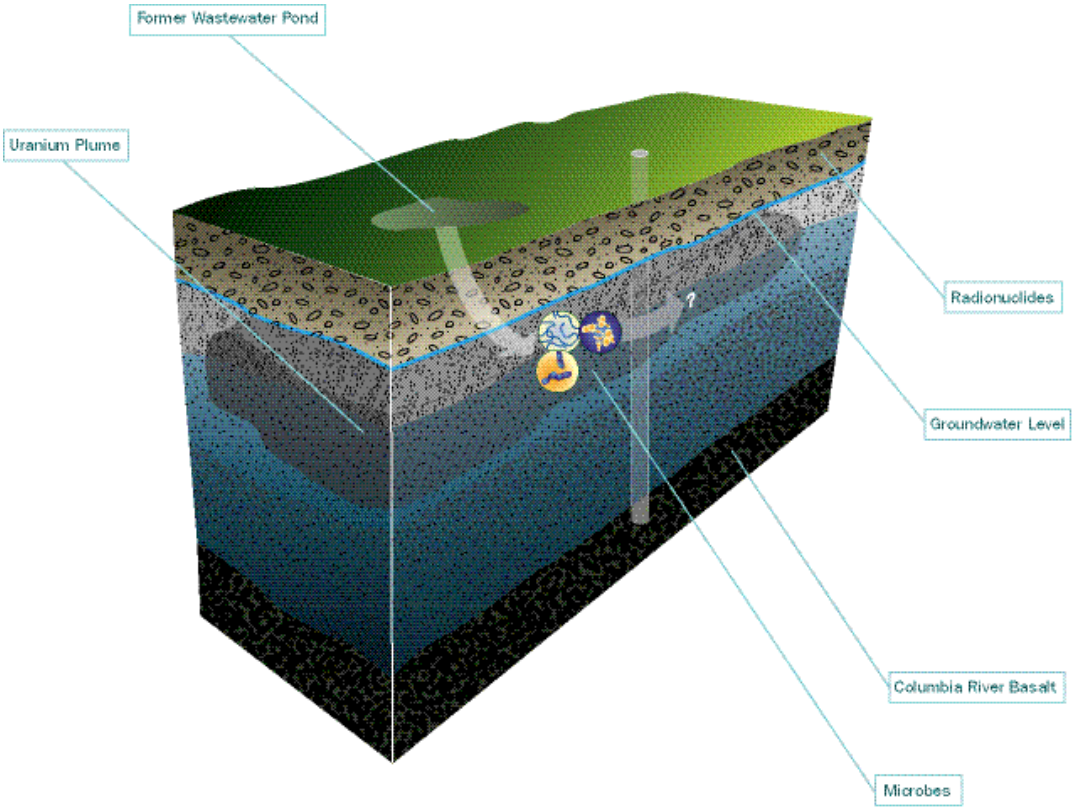
In the future, the DOE JGI will focus on the development of low cost, high quality approaches for finishing microbial genomes. The DOE JGI has become the world leader in producing consistently high quality finished genomes, and with the development of new cost effective finishing approaches will continue to do so in the future for a prioritized group of microbes. To do this, we will develop approaches for reducing cost and improving efficiency of genome finishing, as well as pursuing a broader dialog with the microbial genome community to establish community based quality standards.

Id. Implement high throughput, automated functional studies on all cultured isolates and environments

The DOE JGI will develop methods and partner with other facilities so that the targeted organisms will yield information enabling the advancement of experimental work. In particular, baseline transcriptomics and proteomics, of cells under certain growth conditions will be of great value. Synergistic interactions with other DOE laboratories and facilities providing unique capabilities in areas such as proteomics, bioinformatics, and structural genomics will be sought. These partnerships will provide a springboard for systems-based microbiological investigations.

Ie. Link sequencing pipeline to technologies for isolating organisms from environmental samples

As we move from analysis of individual microbial species to thousands, we must address the major obstacle to the isolation of DNA samples fit for sequencing—that most microbes are difficult, if not currently impossible, to culture in the laboratory. The DOE JGI will target the development of high throughput cost effective technologies for acquiring microbes from environmental samples. These will include single cell targeted cell separation technologies and novel culturing methods, as well as automated DNA isolation approaches so that projects requiring the sequencing of thousands of isolates can be accomplished. A series of platforms and technologies directed at the isolation of genomic DNA from environmental samples will



need to be developed. Accordingly, pre-sequencing sample processing and DNA isolation will represent a major technological focus area for the DOE JGI in the future.

If. Functional analysis of gene products, microbial and metagenomic studies

The DOE JGI has an opportunity to extend its capabilities in the microbial and metagenomic space by adding to them through collaborations with centers with functional screening capabilities such as the DOE Bioenergy Research Centers, as well as the structural genomic pipelines at the DOE light sources among others.

Ig. Align/Integrate with complementary DOE programs

Another response to the changing landscape of microbial and metagenomic sequencing is to carry out projects that are more integrative—drawing researchers and projects together across the DOE complex and programs. In essence, this is the goal of the DOE Genomics: GTL Program, where a few individual species were selected for deep, integrated investigation. Through ongoing discussions with the multiple partners distributed throughout the DOE complex, including various advisory groups and input from the DOE Office of Biological and Environmental Research, the DOE JGI will ensure that its microbial and metagenomic sequencing efforts are aligned as closely as possible to other DOE mission-related studies.

II. Plant Genomics Program Goals

In addition to the microbial genomics and metagenomics program goals, a central goal of the DOE JGI in the next five years will be the sequencing of major bioenergy crops and those essential comparators in order to understand com-

plex traits that can be translated into greater biomass yields, more efficient deconstruction of plant cell walls, improved drought and pest resistance, and easier conversion to potential fuels.

Because of the large and repetitive nature of many plant genomes of DOE-relevance, new sequencing strategies that leverage state-of-the-art technologies will need to be deployed. In addition, comprehensive cataloging of sequence variation, as well as tools for visualizing and analyzing these data will need to be developed.

A deep understanding of the functional adaptations and genetic variation of plants is central to DOE's mission. Rapidly growing plants that require minimal nutrient and water inputs can serve as biofuel feedstocks, primary agents for sequestering carbon in natural and agricultural settings, and active mediators in removing contaminants from soil. The central role of plants in the global carbon cycle, and the susceptibility of this role to biotic and abiotic stress, provides a broader context for plant science in service to the DOE mission. A long-term goal of the DOE JGI is to create a set of genome-based resources and tools that will enable the breeding and/or engineering of plants and plant-associated communities relevant to the DOE mission. Over the next five years we will not only complete the assembly and annotation of nearly a dozen new plant genomes, but we will also create the associated expression and genotyping resources needed to advance our understanding of plant growth and development relative to bioenergy development, carbon cycling, and response to environmental change and phytoremediation.

Plant genomes are very large and complex relative to microbial genomes, attributes that impact the sequencing, assembly, and annotation processes associated with finishing such genomes. As a result, the DOE JGI, with finite resources, must strategically select those plant genomes that are DOE-relevant, informative and obtainable. The new sequencing platforms relax some of the constraints on this selection process but also add new informatics and assembly issues.

The National Research Council panel, charged with assessing the achievements of the National Plant Genome Initiative, noted that "JGI, as the key (in fact, the only) major plant genome sequencing center, is critical to future overall success of the National Plant Genome Initiative (NPGI). JGI's contribution to plant genomics is unique and fundamental, and spans both explicitly energy-oriented projects and projects that broadly inform all of plant biology from evolution through comparative genomics. [We recommend that] DOE continue to take a broad view of DOE JGI's unique position in the plant science community. It is critical to the success of NPGI that JGI continue to serve a broad remit for sequencing and resequencing of plant genomes…"

To attempt to meet this responsibility, we propose several parallel thrusts in plant genomics that take advantage of the DOE JGI's unique ca-

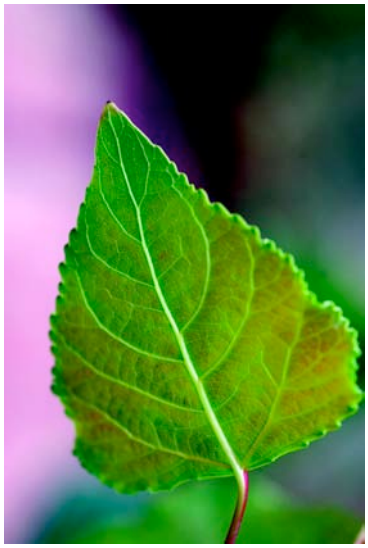
pabilities and scale to drive the development of new biofuel crops, enable reduced greenhouse emissions from agriculture, and mitigate adverse responses to global climate change.

IIa. Sequence: Produce and annotate genome sequences for diverse plant species

High-quality genome sequences provide vital infrastructure for modern plant science. Over the next five years, the DOE JGI will produce annotated genome sequences and transcriptomes for key plants that will not only include species with direct DOE mission relevance (e.g., candidate biomass), but also a diverse set of plants and alga that through comparative analyses will create a framework and methodology for studying DOE-relevant traits (e.g., regulation and composition of lignocellulose biosynthesis, drought tolerance, control of dormancy and reproductive timing, etc.). Since plant genomes are often large and complex, the value of more expensive long-read sequencing technologies must be balanced against the lower cost and higher throughput of new short-read technologies to provide the most useful resources to the research community in a cost-effective manner.

The DOE JGI plant genome sequencing to date has been driven by the Laboratory, Community, and Bioenergy Research Center Sequencing Programs through the proposal of individual species. To promote a coherent and programmatic structure as we go forward, we will establish a panel of advisors from the global plant sciences community whose mission will be to consider white papers for individual species and to select and prioritize plant genomes for sequencing by the DOE JGI. Criteria will include DOE mission relevance, the existence of active research communities, and the broader impact on plant science. In addition, we will continue to develop and support "flagship" genomes. While the cost of producing "rough draft" quality genome sequences will continue to decline over the next five years, complete, accurate, and well-annotated "reference" genomes will likely remain costly, and can only be tackled by a dedicated facility with a long-term focus on genomics. High-quality reference sequences are essential both as a platform for research in the organism in question, but also as anchors for comparative studies. We propose that over the next five years, at least a half-dozen plant genomes will be designated as "flagship" genomes that will be the subject of targeted finishing and deep transcriptome sequencing to enable accurate and complete annotation. All of these have major DOE mission relevance particularly for bioenergy. These include:

- ***Populus trichocarpa,*** 400 MB, a rapidly growing model tree and potential woody biomass crop.
- ***Sorghum bicolor,*** 700 MB, the second largest potential biofuel crop in the U.S. (after its close relative, maize) and a model for other C4 grasses.
- ***Glycine max*** (soybean), 980 MB, a prominent agricultural crop, source of biodiesel, and model for the nitro-gen-fixing legumes.
- **Foxtail Millet**, 400 MB, a grass, model for switchgrass, a potential cellulosic biomass crop.
- ***Chlamydomonas reinhardtii***, 110 MB, algal model organism, the preeminent model green alga and focal point of DOE algal research in photosynthesis.
- ***Physcomitrella patens***, 480 MB, moss, a tractable model organism for studies of the cell wall, the principal component of terrestrial biomass and a key target for processing to sugars for bioenergy.
- Two-to-four additional genomes to be selected by the DOE JGI plant genome advisory panel.

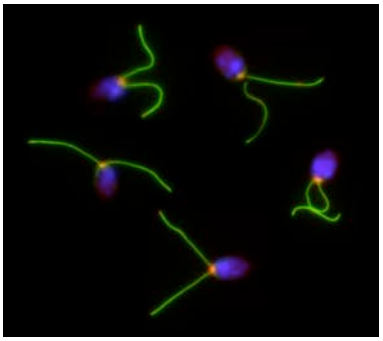


Populus trichocarpa (poplar)

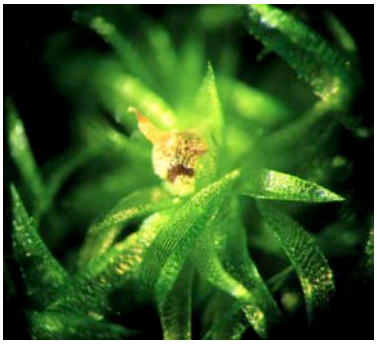
Sorghum bicolor

Glycine max (soybean)

Setaria italica (Foxtail Millet)



Chlamydomonas reinhardtii



Physcomitrella patens

Finally, we must develop cost-effective approaches to sequencing complex genomes. Some economically important plants have extraordinarily complex genomes that are 5-6 times larger than any plant or animal genome tackled to date.

For example, loblolly pine (~20 billion bases [Gb]) and hexaploid wheat (~16 Gb) have been identified as national priorities by the Interagency Working Group on Plant Genomics, and are the focus of coordinated academic, governmental, and industrial research groups around the world.

Novel strategies will need to be developed to tackle these genomes and to take maximal advantage of new higher throughput approaches. The DOE JGI is uniquely positioned to undertake an extended effort to develop genomic resources for these species through innovative applications of new and old sequencing technologies and computational methods. This will require significant research in sequence assembly.

IIb. Function: Develop resources for comparative and functional genomics in key clades

When the human genome sequence was completed, the gap between sequence and function was accentuated. This realization spurred the U.S. National Institutes of Health to initiate coordinated efforts to characterize the function of sequence elements in the human genome. As a parallel to this human-genome-centric project, the DOE JGI will develop resources for comparative and functional genomics in key clades. The long-term aim of this thrust is to identify the functional genetic elements in plant genomes by sequencing related species at various phylogenetic distances for comparative purposes, as well as to develop datasets and bioinformatic tools. This effort can serve as a focal point for the generation of large complementary datasets from other DOE laboratories as well as the larger research community.

Over the next five years, the DOE JGI will develop both deep transcriptome and comparative sequence datasets centered on *Arabidopsis*, *Populus,* and the clade of C4 grasses that includes switchgrass, *Miscanthus*, and sugarcane. These two focal points provide programmatic coherence to existing projects in the DOE JGI sequencing queue and contribute to DOE Bioenergy Research Center projects and infrastructure.

In addition, during this time, we anticipate laying the scientific groundwork at the DOE JGI to develop some resources for functional genomics in plants. As this thrust develops, we anticipate that capabilities at other National Laboratories (for example, proteomics efforts at PNNL and elsewhere) can be brought to bear on exploring proteins and protein complexes, along with the development of gene expression profiles of flagship plant species under different circumstances or treatments.

IIc. Characterize plant genetic diversity

Natural genetic variation provides an essential resource for the improvement of bioenergy crops through breeding, and provides an avenue for understanding the response of plant populations to climate change. The long-term goal of this third thrust is to gain knowledge of the genetic diversity contained within individual plant species and to be able to predict a plant's responses to alternating environmental parameters. This understanding could be applied to the domestication of energy crops, the strategic deployment of phyto-remediation systems, and predictive modeling of entire ecosystems. The DOE JGI will take advantage of high-throughput resequencing technologies to characterize natural genomic (sequence) variation in plants and to facilitate the association of such variation with biofuel sustainability and adaptation to climate change in conjunction with the DOE JGI partner labs.

Natural genetic and epigenetic variation in plant systems can be at once ubiquitous and inconspicuous, or even hidden. Ongoing resequencing projects in maize, *Arabidopsis*, *Populus*, and *Mimulus* will yield tools and analytical approaches that will foster high-throughput characterization of within-species diversity and allow breeding for highly

specialized phenotypes, such as those adapted to extreme environments. These technologies can piggyback on tools, such as those that emerged from the International HapMap Project focused on the nature of human genetic variation. In addition to describing the sequence diversity of plants we must also consider their interaction with associated microbial communities, which makes important contributions to carbon sequestration as well as host success. The characterization of these communities and their relationships to host genotype, soil chemistry, and agricultural practices is an essential component of the DOE JGI plant genome program, and has notable synergies with other DOE JGI metagenomics efforts.

III. Informatics and Computational Biology

Genome sequence is initially transformed from raw data into useful biological knowledge through computation. These analyses include the *de novo* assembly of genomes, identification of conserved and variant sequences, annotation of genes and other functional elements, and comparative analyses across multiple genomic datasets that illuminate populations and ecosystems. Computational biology is embedded within the scientific programs of the DOE JGI, and informatics associated with the management and processing of sequence data is an integral part of the production sequencing effort. As sequencing capacity increases informatics capabilities must also grow, lest they become a bottleneck that could limit new applications of genomics to problems of DOE relevance. An essential component of the DOE JGI Strategic Plan is therefore the continued development of systems to capture, manage, and interpret increasing volumes of genomic data.

The extensive computational expertise and infrastructure of the DOE National Laboratory system provides a unique advantage of scale to DOE JGI, far beyond that found in smaller genome-sequencing operations. Leveraging these resources promises to provide DOE JGI users with unique computational capabilities paralleling the unique high-throughput sequence generation. Given the trajectory of sequencing at the DOE JGI, both operational and programmatic areas will require significant scaling of analytical capabilities, development of new computational capabilities, which will entail an ongoing dialog to coordinate access to computational resources. The path forward requires the following efforts:

IIIa. Develop data management capabilities associated with the exponential growth of sequencing and increasing complexity of DOE JGI projects

The ever-increasing throughput of genomic technologies presents challenges for the management and distribution of raw and processed genomic datasets at DOE JGI. In the areas of metagenomics, comparative genomics, and population genomics, terabase (10^{12} bp)-scale "grand challenge" projects are already emerging, and will become the norm. DOE JGI's production informatics capabilities must keep pace with the generation of these raw datasets, working hand-in-hand with the technology development and production sequencing teams to ensure that appropriate data management capacities and capabilities are matched with new sequencing instrumentation and methodologies. We anticipate that this will require the development and deployment of advanced computing clusters and associated data management software tools to organize data internally and to make slices of these data readily available to the scientific community, and to over come such bottlenecks as the transfer of large blocks of data in and out of memory for analysis and use. Novel developments in sequencing will likely need to be paired with new developments in computation as the respective "Moore's Laws" of the two fields push genome informatics forward.

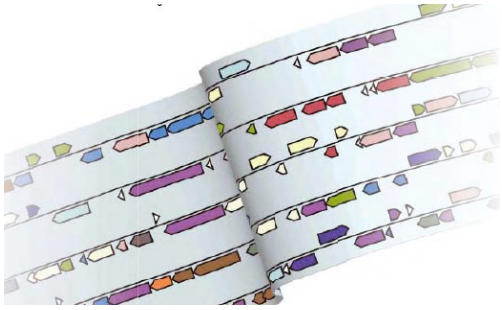
IIIb. Lead the development of metadata and sequence quality standards

While raw sequence is readily described in standard formats, the increasing diversity of DNA and RNA sources requires the development of an appropriately rich set of standards for representing the "metadata" that captures the relevant variables associated with each sample. These metadata are vital for deep comparative analyses of metagenomic, population resequencing, transcriptomic, and other datasets. The diversity of projects coming out of ongoing GTL consortia, the DOE Bioenergy Research Centers, and the emerging Systems Biology Knowledgebase puts DOE JGI at the forefront of the development of metadata standards. Similarly, the explosion of plant genome sequence and variation data will drive the establishment of data standards for plant science, with DOE JGI as a focal point. The high-throughput generation of microbial genomes associated with the GEBA project will require the development of new standards for sequence quality, which will need to be tailored to the strengths and weaknesses of emerging new sequencing technologies. Thus DOE JGI scientists and informaticists must continue to play leading roles in the development of new data standards for sequence-based science as they have been with the Genomic Standards Consortium (GSC) to capture a more information-rich collection of complete genomes and metagenomic datasets.

IIIc. Harness the computational expertise and infrastructure of the DOE National Laboratories for advanced genome analysis

The DOE National Laboratory system has a long and illustrious history of developing and applying high performance computing for the physical sciences, including peta-scale data storage, management and retrieval, and extensive grid, cluster and supercomputing resources. To meet DOE JGI's computing needs in the face of exponential ge-

nomic data growth will require harnessing these capabilities for high throughput biological sequence analysis where appropriate as determined by the science. Within the next two years, DOE JGI will work with Laboratory partners to develop scalable software infrastructures for large-scale



sequence comparisons (e.g., BLAST), profile-based methods (e.g., HMMer), phylogenetic analysis, and comparative genomics. Initial meetings to identify and define computational challenges faced by the DOE JGI and foster collaborative efforts are already underway. Through these developments, the DOE National Laboratories will expand the capabilities of DOE JGI and its users by enabling large-scale computations that cannot be carried out on small scale computing clusters. We anticipate that these developments will create positive feedback; as more complex large-scale computational analyses become feasible, new and larger-scale applications of sequencing will be envisioned by DOE JGI and its users, requiring, in turn, even larger computational analyses.

IIId. Develop and apply new algorithms for high throughput sequence analysis

The exponential growth of sequencing and the parallel growth of computing resources requires the development and application of appropriately high throughput algorithms. Obvious areas for advancement include large genome assembly, assembly of data from multiple sequencing platforms, metagenomic analysis, annotation of protein-coding and non-coding RNA genes, phylogenetic reconstruction, quantitative and population genetics, large-scale comparative genomics, and metabolic analysis. The identification, application, and when necessary development of these algorithms at DOE JGI will be driven by the JGI scientific programs. In many cases, relevant algorithms and tools can and should be adopted or adapted from the existing toolkit of the biomedical genomics community. In other cases, such as the analysis of complex environmental metagenome samples or the assembly of the most complex plant genomes, DOE JGI may face problems not yet tackled by biomedical researchers. Indeed, we anticipate that the transition from single-genomes to complex terabase-scale projects will inevitably raise unique and larger scale computational problems requiring custom solutions that will combine existing bioinformatics methods with novel algorithms targeting advanced computational platforms. The assemblage of robust teams of DOE JGI scientists who can collaborate with users to address these problems is essential to the success of the DOE JGI scientific and user programs.

IIIe. Develop state-of-the-art comparative genome data portals for microbes and plants

The DOE JGI has developed robust comparative data portals for microbes (the Integrated Microbial Genomes [IMG] system) and plants (phytozome.net), and continues to support diverse eukaryotic genomes on the DOE JGI genome browser. The continued development and diversification of these data portals is essential to the strategic vision of the DOE JGI as a focal point of sequence-based science. The systems have been created by the corresponding DOE JGI scientific programs as an integral part of their scientific missions, and serve as a community resource for comparative analysis and annotation of diverse genomes in an appropriately integrated

context. These portals provide a unified, "one stop shop" for genomic data, enabling users to navigate across genomes, gene repertoires, and pathways; to retrieve relevant data slices; to perform increasingly complex and customized queries and analyses over the web; and to integrate open source software with DOE JGI-developed comparative algorithms. Close ties to users fostered by linkages to other data sites and regular user workshops will ensure that these systems remain at the center of their respective research communities, and that they contain up-to-date, sequence-based and functional datasets with associated analyses.

IV. JGI Users Programs, and Value Added Technologies

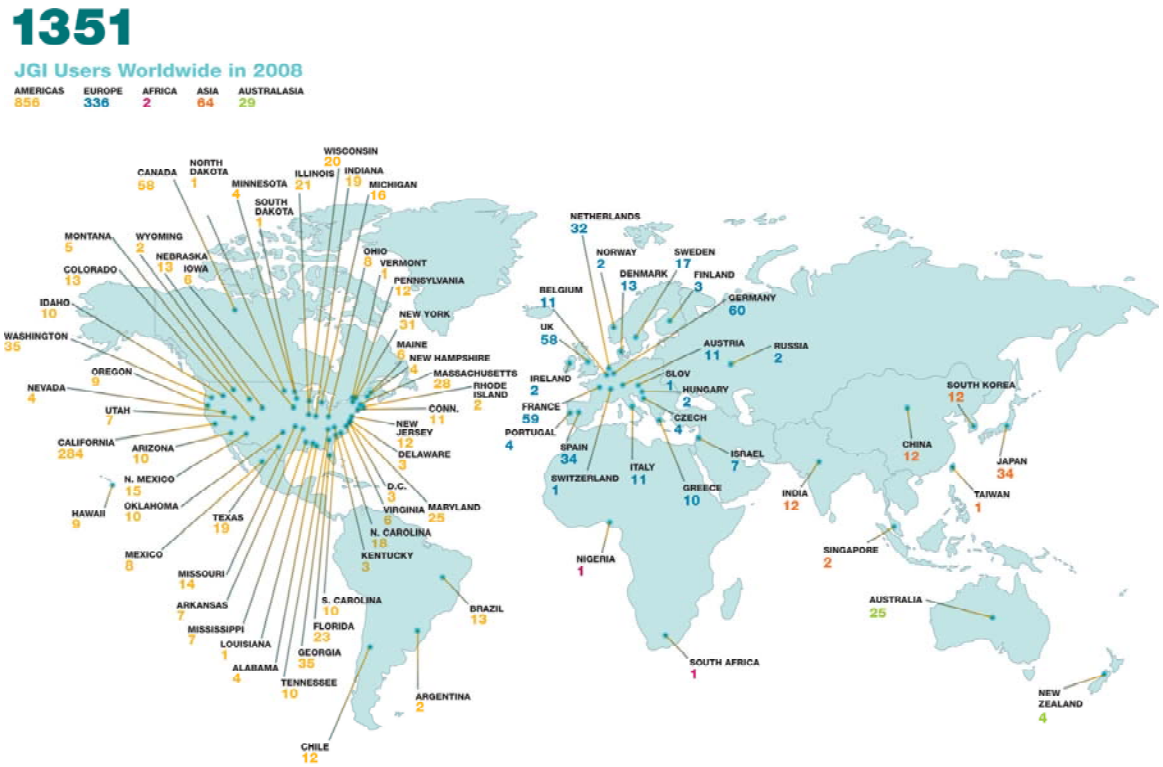
The enormous mission-relevant impact of the DOE JGI's science that has occurred in the last five years is largely the result of its user programs. The user programs began with the concept that investigators would propose genome sequencing of an organism or a small number of organisms of interest to a community of investigators studying a DOE-relevant problem. Success with user projects often relies on prior experience with new technologies developed or adapted by a small core of DOE JGI scientists carrying out groundbreaking sequenced-based biology. Examples include metagenome assembly, single-cell sorting and genome amplification, and tools development for the comparison of microbial genomes. In each case, new strategies and techniques were developed when users brought challenging, DOE-relevant projects to the DOE JGI that were tackled by a collaborative team of external users and DOE JGI scientists. This model will continue in the future and the DOE JGI will continue to slowly grow the critical core of internal scientists dedicated to carrying out their own sequence based science as well as working with users to expand the reach of genomics into biofuels production, global carbon cycling and biogeochemistry.

With the development of routine procedures for shotgun sequencing and genome assembly and the introduction of massively-parallel new technology-based sequencing platforms, DOE JGI user proposals frequently request the sequencing of multiple organisms or environments that allow genome comparisons on a much larger scale than the past. In the future, the DOE JGI's unique contribution will increasingly be to work with users in tackling projects of significant scale and depth that are beyond the capabilities of smaller facilities. As described in the program plans above, we anticipate carrying out sequencing projects that entail sequencing hundreds of microorganisms, dozens of samples for deep metagenome sequencing, or hundreds of plant strains to understand their variation. The DOE JGI will facilitate these large-scale efforts by placing particular emphasis on automating aspects of sample collection and pre-sequencing molecular biology, areas that still remain rate-limiting for most genome projects. Finally, DOE JGI will expand its analytical capabilities to be able to handle extremely large datasets while developing user-friendly tools that its user community can exploit to answer important questions pertaining to energy and the environment. These laboratory and computer based capabilities will be made available to users to facilitate ground-breaking DOE relevant science.

IVa. Grand Challenge Projects

From its inception as a user facility, the DOE JGI has enabled investigators to address large-scale genomic problems unique to the scale of the Institute's capacity for production of sequence and its analysis. While presently the average user project requires producing a few million to hundreds of millions of base pairs, in the near future we will be helping users to address problems requiring may orders of magnitude more sequence as well as various forms of "meta-data." For example, metagenome studies to date have been largely used to construct metabolic snapshots of particular environments. In the future, we will be taking highly granulated high resolution photographs of microbial communities over time and space in order to capture how a community responds to stresses such as drought, heat, elevated CO_2 or salt concentrations, and other experimental interventions. To do this will require both deeper and broader sequencing, deployment of new strategies for managing these large datasets and the development of new tools for sequence assembly and analysis. Success in these endeavors will lead to a much more complete understanding of how nutrients and toxins move through DOE-relevant environments and how low abundance organisms contribute essential activities in plant health, biomass degradation and handling of environmental toxins.

The DOE JGI's current microbial genome projects focus on the sequencing of one or a small number of organisms. It is expected that future projects will entail sequencing of hundreds of organisms, perhaps isolated as single cells without cultivation, in order to deeply probe community structure and population diversity in DOE relevant environments. To accomplish such goals will require the development of capabilities to handle large numbers of samples, sequence single cells, and accurately compare their genomes. Success in these efforts will be measured by filling in the microbial tree of life with sequenced organisms that will in turn lead to discovery of enzymatic activities that can be deployed in biomass degradation, biofuels synthesis, or degradation/sequestration of toxins, and to improved functional annotation of micro-



bial communities obtained from DOE relevant environments.

Likewise, plant genome projects currently target bioenergy feedstocks and important comparators. Once these reference genomes are in hand, new technologies will be exploited to examine the genomes of hundreds of variant strains carrying traits of DOE interest in order to identify the genes responsible for these characteristics. Success will be marked by better understanding of plant biomass productivity, cell wall metabolism, and resistance to drought and pests.

Finally we expect to solicit from users large-scale projects that examine DOE-relevant crosscutting themes, such as the molecular basis of plant-microbial interactions that will require combined sequencing and analysis of both the plant and the microbial consortia associated with its roots, stems and leaves. These projects represent just a few of the grand challenge calls that the DOE JGI will issue to enable users to solve important problems in energy and the environment by exploiting the DOE JGI's unique capability in genomics of scale. As has been our history, it is expected that DOE JGI scientists will collaborate with community-based users in many of these large-scale projects to maximize the scientific value obtained from expenditure of DOE JGI resources.

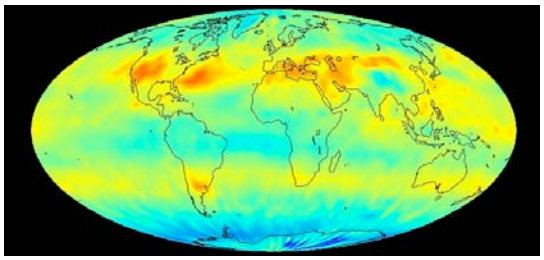
IVb: Interacting with other DOE user facilities and research centers

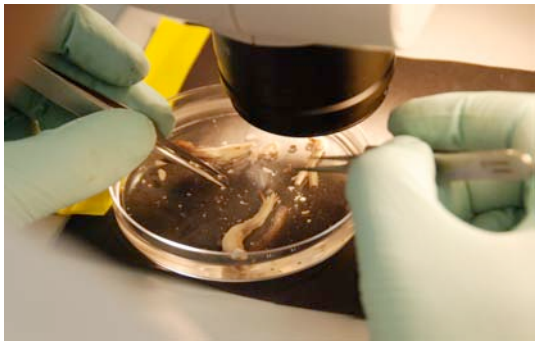
In addition to the DOE JGI, DOE operates other user facilities, each with unique capabilities. In the future we plan to combine high throughput genomics with capabilities of other DOE user facilities to offer systems based approaches for studying DOE relevant problems that may be unattainable through genomics alone. As an example of this, the DOE JGI is exploring the development of a partnership with PNNL's Environmental Molecular Sciences Laboratory (EMSL), to link genomics and proteomics by calling for proposals that can synergistically exploit the two facilities' combined capabilities. The goal of such projects would not be simply to create large projects, but rather to seek particular examples from the user community of how EMSL facilities and capabilities can synergistically add value to the DOE JGI's sequencing and analysis projects. The DOE JGI envisions interactions with other DOE user facilities that would exploit High Performance Computing centers, to solve particular analytical problems associated with large datasets, structural genomic projects with synchrotron radiation lightsources, or engineering of biological environments with Nanoscale Science Research Centers, to name just a few.

The DOE Bioenergy Research Centers and similar BER facilities planned for the future embody a new set of priority users for the DOE JGI. These facilities represent opportunities to collaborate with highly focused investigators in developing novel capabilities that can then be offered to DOE JGI's broader user community. For these collaborations to be successful, rapid planning, technological flexibility and a project management staff that has expert knowledge in DOE JGI's sequencing and analysis capabilities are required.

IVc. Proactive project management

As the number of organisms and types of projects approved for sequencing has expanded, the DOE JGI has assembled a committed group of formally trained project managers and developed internal tools for tracking project progress and communication with users. This system must be completely automated and continuously refined to mirror DOE JGI's evolving processes so that all project stakeholders can immediately find the status of any project and begin to solve problems as soon as they arise. By doing this, the DOE JGI will be uniquely positioned to carry out multiple projects that each require processing of hundreds of sam-





ples. This in turn will facilitate genome science of a scale that that does not currently exist anywhere. In order to do this, the DOE JGI must build sample handling and project tracking systems that are fully integrated with its laboratory information and data management systems and that include user interfaces that streamline sample submission, project tracking and data delivery.

IVd: Providing pre- and post-sequencing value-added capabilities to users

Historically, the DOE JGI as a user facility has primarily provided users with access to high throughput sequence data and analysis capabilities. Going forward, the DOE JGI user programs will emphasize value-added pre- and post-sequencing capabilities. Pre-sequencing capabilities will include various technologies for collecting and processing large numbers of samples, ranging from single cell isolation and amplification to rRNA removal for metagenome transcriptome profiling and DNA capture for analysis of very large genomes such as loblolly pine. While most interactions with users in the past have been via DNA samples received in the mail, users may come to the DOE JGI in the future to access the pre-sequencing technologies required to process their samples, combining their organism-specific experience with the substantial technical experience of the DOE JGI scientists and staff. Various technologies will be developed and or adapted at the DOE JGI to provide users with state of the art approaches for preparing samples destined for sequencing.

Just as value-added pre-sequencing technologies will be expanded, increasing emphasis will also be placed on providing expanded post-sequencing analysis capabilities to users. Though the ability to generate sequence is dramatically increasing, the ability to analyze sequence represents an ever-tightening bottleneck. The DOE JGI will place an increased emphasis on exploiting the high performance computing capabilities and informatics expertise within the DOE National Laboratory system to provide users with innovative and robust tools that they can deploy to convert sequence data into knowledge. DOE JGI is positioning itself to play a leadership role in this domain through investment in IT infrastructure, generation of large datasets, and collaborations with DOE high performance computing centers.

IVe: Next generation sequencing technologies

A major continued strategic theme will be to remain state of the art in DNA sequencing capabilities and associated technologies. With the massive financial incentive for private industry to develop technologies to enable personal genome sequencing, next generation sequencing technologies are almost exclusively being developed in the commercial sector. Thus it is expected that the DOE JGI will remain an early adopter of new sequencing platforms as they become available. Through early access, the DOE JGI will be able to assess nascent technologies for their potential utility to DOE JGI users. The DOE JGI will also continue to focus on maintaining our alpha/beta

user relationship with existing DNA sequencing vendors. Through this the DOE JGI is in position to benefit from early access to better performing sequencing reagents, new kits for specialized library construction, and software upgrades to increase sequencing throughput and accuracy.

While the DOE JGI will not be creating the next generation of DNA sequencing machines we will be adapting these instruments and associated technologies to meet the needs of the Institute and its users. The DOE JGI's capabilities in automation and high throughput liquid handling will be employed for automating aspects of sequencing associated molecular biology and sample preparation activities that serve as a bottleneck in scaling next generation sequencing. In addition, a major focus will be in reducing ergonomic challenges of next generation sequencers in a production environment.

Summary

We expect that over the next five years, the DOE JGI will build upon many of the capabilities and activities it offers now to users by developing a new set of value-added resources that address the imminent grand challenges in genomics. To maintain its currency as a user facility focused on tera-scale genomics relevant to DOE missions, it will foster the development of both direct and indirect groups of users. The direct users will initiate projects of scale and complexity uniquely suited to the DOE JGI's capabilities, while the indirect users will be represented by even greater communities of investigators that will draw upon the encyclopedic plant, microbial and metagenomic datasets systematically generated by the DOE JGI. The Institute will continue as one of the global leaders in generating sequence while increasingly contributing to the development of robust and scalable systems for the management, analysis, and distribution of genomic data generated by the DOE JGI and elsewhere.

The activities proposed in this planning document will be essential for advancing the frontiers of bioenergy, carbon cycling and biogeochemistry. The goals outlined are aligned with the Institute's core capabilities, ongoing technological advances in the field, and the central DOE mission needs. With this "living" Strategic Plan, we begin what will be an ongoing conversation to further define the specific targets, strategies, organizational structure and resources needed to accomplish these goals.

Appendix

A1. DOE JGI Five-Year Planning Group

External Participants

Adam Arkin (UC Berkeley) *New Tech*, Steve Beckwith (UCOP, Vice President of Research) *Astronomy* (Hubble Telescope Group Lead), Randy Berka (Novozymes*) Microbial*, Jeff Dangl (University of North Carolina) *Plants*, Joe Ecker (Salk Institute) *Plants*, Drew Endy (Stanford) *New Tech*, Mike Himmel (NREL) *Computing*, Rob Phillips (CalTech) *New Tech,* Steve Quake (Stanford) *New Tech*, Karin Remington (NIH) *Computing*, James Siegrist (UCB Physics) *High Energy Physics*, John Taylor (UCB) *Microbial/ Fungi*, Jim Tiedje (Michigan State) *Microbial*.

Bioenergy Research Centers

Jay Keasling (UCB/BRC) *New Tech*, Martin Keller (ORNL/BRC) *Microbes*, David Mead (Lucigen/Great Lakes BRC) *Microbes.*

DOE JGI Partner Lab Participants

Scott Baker (PNNL*) Microbial*, Jim Bristow (DOE JGI) *User Facility/Ops*, Patrick Chain (LLNL) *Microbial and Meta*, Bob Cottingham (ORNL) *Computing/Informatics*, Chris Detter (LANL) *Microbial and Meta*, Jonathan Eisen (UC Davis) *Microbial*, Peg Folta (LLNL) *Computing*, David Gilbert (DOE JGI) *Floater,* Yakov Golder (DOE JGI) *Computing*, Susan Lucas (DOE JGI) *New Tech*, Reinhold Mann

(DOE JGI) *New Tech*, Reinhold Mann (ORNL) *Computing/Informatics*, Rick Myers (HudonAlpha) *New Tech*, Len Pennacchio (DOE JGI) *New Tech,* Doug Ray (PNNL) *User Facility/Ops*, Gary Resnick (LANL) *User Facility/Ops*, Dan Rokhsar (DOE JGI) *Plants*, Margie Romine (PNNL) *Microbial*, Nina Rosenberg (LLNL) *User Facility/Ops*, Eddy Rubin, Director (JGI) *Floater*, Jeremy Schmutz (HudsonAlpha) *Plants*, Tom Slezak (LLNL) *Computing*, Jerry Tuskan (ORNL) *Plants*, Ray Turner (DOE JGI) *User Facility/Ops*.

DOE Headquarters

Dan Drell (DOE) *Floater*, Susan Gregurick (DOE) *Computing,* Marvin Stodolsky (DOE) *New Tech*.

A2. Safety

The DOE JGI is fully committed to providing a safe and healthful workplace to prevent work-related injuries and illnesses. At the DOE JGI, Integrated Safety Management (ISM) is the method used to ensure that all operations are planned and executed to protect the health and safety of each employee, the public, and the environment. Many changes and improvements were made to the internal safety program over the past couple of years in response to new Environmental, Health, and Safety (EH&S) requirements of DOE. These improvements include an increase in safety staff, development of new processes for early intervention and injury case management specific to ergonomic issues, and the development of a system for responding to ergonomic issues. The DOE JGI has created and launched a forward-looking program, "Beyond HSS" (Health, Safety and Security), that focuses on various areas of longer-term improvements in the area of Laboratory Safety and Issues Management. In the future a high priority will be placed on seeing that progress is made in these areas to further ensure a safety environment where the DOE JGI staff can achieve excellence in science.